C-Index (Concordance)

General Problem Setting

- $(X, Y) \sim \wp$, where:
  - Y is a continuous random variable
  - X is a random vector
  - $\wp$ is their unknown joint distribution
- M(X) prediction model: X $\mapsto$ [Prediction of Y]

A. No Censored Data

- Dataset:

$$\begin{pmatrix} X_1 & Y_1 \\ . & . \\ . & . \\ . & . \\ X_N & Y_N \end{pmatrix}$$

Validation Metric:

- $(X_A, Y_A) \sim \wp$, $(X_B, Y_B) \sim \wp$ and they are independent.
- $Z(M(X), \wp) = P(M(X_A) < M(X_B) | Y_A > Y_B)$
- Estimate of $Z(M(X), \wp)$:

$$Z(M(X), \widehat{P}) = \frac{\sum_{i,j} 1(M(X_i) > M(X_j), Y_i < Y_j)}{n(n-1)}$$

where $\widehat{P}$ is the empirical distribution of the data.

B. Censored Data - Fixed censoring time

- τ: fixed censoring time. All the observations are censored at the same time.
- Dataset:

$$\begin{pmatrix} X_1 & Y_1 \\ . & . \\ X_i & Y_i^+ \\ . & . \\ X_N & Y_N \end{pmatrix}$$

where Y_i^+ denotes that the event doesn't occur before the censoring time.

Validation Metric:

- $(X_A, Y_A) \sim \wp$, $(X_B, Y_B) \sim \wp$ and they are independent copies.
- $Z(M(X), \wp) = P(M(X_A) > M(X_B) | Y_A < Y_B, Y_A < \tau)$
- Estimate of $Z(M(X), \wp)$:

$$\frac{\sum_{i,j} 1(M(X_i) > M(X_j), Y_i < Y_j, Y_i < \tau)}{\sum_{i,j} 1(Y_i < Y_j, Y_i < \tau)}.$$

C. Censored Data - Distribution of right censoring times

- C right censoring time
- $C \sim G$
- Dataset

$$\begin{pmatrix} X_1 & Y_1 & \Delta_1 \\ . & . & \\ X_i & Y_i^+ & \Delta_i \\ . & . & \\ X_N & Y_N & \Delta_N \end{pmatrix}$$

where Δ_i is equal to 0 if the observation is censored, 1 otherwise.

- Non-informative (random) censoring
  - Each subject has a censoring time that is statistically independent of their failure time.
  - The observed value for an individual is the $\min(C_i, Y_i)$

Validation Metric:

- $(X_A, Y_A) \sim \wp$, $(X_B, Y_B) \sim \wp$ and they are independent
- $Z(M(X), \wp) = P(M(X_A) > M(X_B) | Y_A < Y_B, Y_A < \tau)$
- Estimate of $Z(M(X), \wp)$:

$$\frac{\sum_{i,j} \Delta_i (\widehat{G}(Y_i))^{-2} 1(M(x_i) > M(x_j) 1(Y_i < Y_j) 1(Y_i < C)}{\sum_{i,j} \Delta_i (\widehat{G}(Y_i))^{-2} 1(Y_i < Y_j, Y_i < C)}$$

where $\widehat{G}$ is the estimated censoring distribution obtained through standard Kaplan-Meier estimation.

Example

The following data come from Klein and Moeschberger (1997) *Survival Analysis Techniques for Censored and truncated data*, Springer. National Longitudinal Survey of Youth Handbook The Ohio State University, 1995. The descriptions of the variables are below.

- `group` Disease Group 1-ALL, 2-AML Low Risk, 3-AML High Risk
- `t1` Time To Death Or On Study Time
- `t2` Disease Free Survival Time (Time To Relapse, Death Or End Of Study)
- `d1` Death Indicator 1-Dead 0-Alive
- `d2` Relapse Indicator 1-Relapsed, 0-Disease Free
- `d3` Disease Free Survival Indicator 1-Dead Or Relapsed, 0-Alive Disease Free)
- `ta` Time To Acute Graft-Versus-Host Disease
- `da` Acute GVHD Indicator 1-Developed Acute GVHD 0-Never Developed Acute GVHD)
- `tc` Time To Chronic Graft-Versus-Host Disease
- `dc` Chronic GVHD Indicator 1-Developed Chronic GVHD 0-Never Developed Chronic GVHD
- `tp` Time To Chronic Graft-Versus-Host Disease
- `dp` Platelet Recovery Indicator 1-Platelets Returned To Normal, 0-Platelets Never Returned to Normal
- `z1` Patient Age In Years
- `z2` Donor Age In Years
- `z3` Patient Sex: 1-Male, 0-Female
- `z4` Donor Sex: 1-Male, 0-Female
- `z5` Patient CMV Status: 1-CMV Positive, 0-CMV Negative
- `z6` Donor CMV Status: 1-CMV Positive, 0-CMV Negative
- `z7` Waiting Time to Transplant In Days
- `z8` FAB: 1-FAB Grade 4 Or 5 and AML, 0-Otherwise
- `z9` Hospital: 1-The Ohio State University, 2-Alferd , 3-St. Vincent, 4-Hahnemann
- `z10` MTX Used as a Graft-Versus-Host- Prophylactic: 1-Yes 0-No

Let's start by loading the following packages and data:

```r
library(survival)
library(KMsurv)
library(survAUC)
library(dynpred)
data(bmt)
```

Let's look at the data:

```r
attach(bmt)
head(bmt)
```

```
##   group   t1   t2 d1 d2 d3   ta da  tc dc tp dp z1 z2 z3 z4 z5 z6   z7 z8
## 1     1 2081 2081  0  0  0   67  1 121  1 13  1 26 33  1  0  1  1   98  0
## 2     1 1602 1602  0  0  0 1602  0 139  1 18  1 21 37  1  1  0  0 1720  0
## 3     1 1496 1496  0  0  0 1496  0 307  1 12  1 26 35  1  1  1  0  127  0
## 4     1 1462 1462  0  0  0   70  1  95  1 13  1 17 21  0  1  0  0  168  0
## 5     1 1433 1433  0  0  0 1433  0 236  1 12  1 32 36  1  1  1  1   93  0
## 6     1 1377 1377  0  0  0 1377  0 123  1 12  1 22 31  1  1  1  1 2187  0
##   z9 z10
## 1  1   0
## 2  1   0
## 3  1   0
## 4  1   0
## 5  1   0
## 6  1   0
```

```r
table(group)
```

```
## group
##  1  2  3
## 38 54 45
```

```r
summary(t2[group == 1])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   123.8   400.5   609.4  1153.0  2081.0
```

```r
summary(t2[group == 2])
```

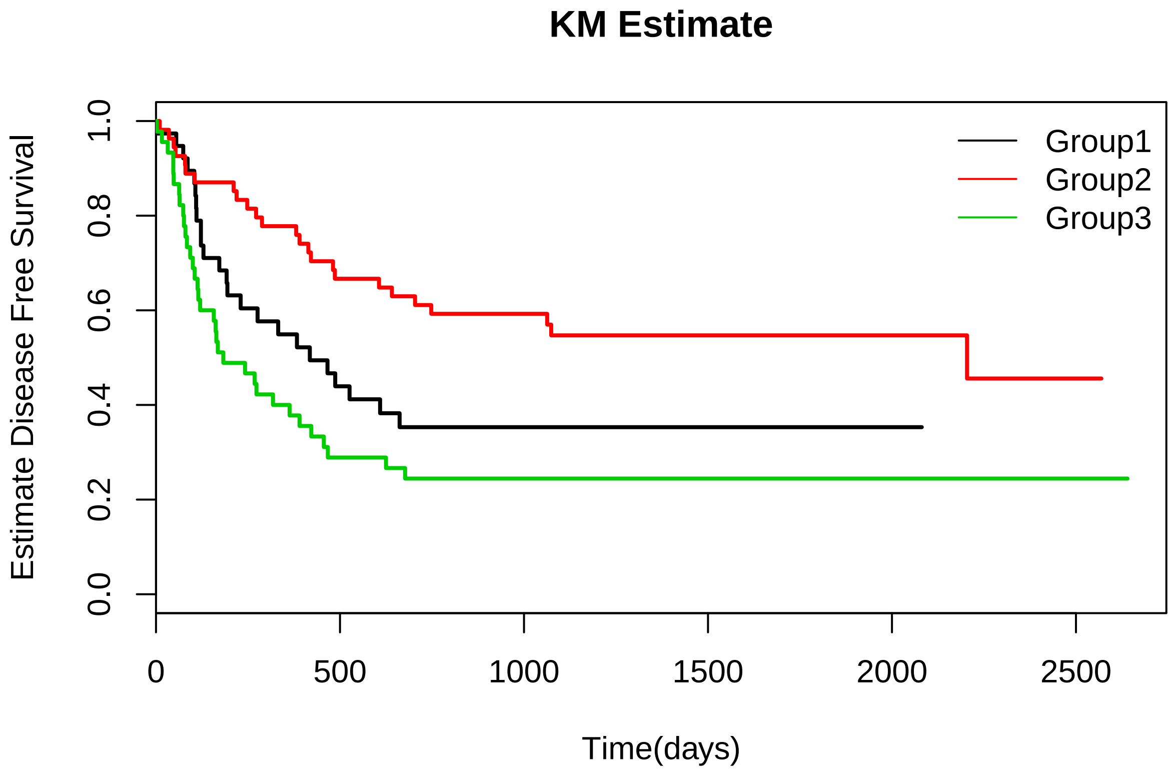
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.0   396.0   993.5  1065.8  1647.5  2569.0
```

```r
summary(t2[group == 3])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0    84.0   183.0   587.3   677.0  2640.0
```

Let's estimate the KM disease-free survival curves for the three groups:

```r
mod.surv = survfit(Surv(t2,d3) ~ group, data=bmt )
plot(mod.surv, ylab="Estimate Disease Free Survival", xlab="Time(days)",
     main="KM Estimate", col=1:3, lwd=2)
legend("topright", legend=c("Group1", "Group2", "Group3"), col=1:3 , lty=1, bty="n")
```



Let's consider Group 3 only and fit a Cox proportional hazards model using `t2` as the event time and `d3` as the censoring information with a single predictor `z1` (Patient Age in Years).

```r
# Split group 3 into training and testing sets
bmt.train = bmt[93:114,]
bmt.test = bmt[115:136,]

# Fit a Cox proportional hazards regression model using the training data
mod.surv.train = coxph(Surv(t2, d3) ~ z1, data = bmt.train)

# Compute prediction summaries for the remaining validation component of the data set
lpnew = predict(mod.surv.train, new.data = bmt.test )

Surv.rsp = Surv(bmt.train$t2, bmt.train$d3)    # The outcomes of the training data
Surv.rsp.new = Surv(bmt.test$t2, bmt.test$d3)  # The outcome of the test data
```

We will use Uno's estimator which is based on inverse-probability-of-censoring weights and does not assume a specific working model for deriving the predictor lpnew. It is assumed, however, that there is a one-to-one relationship between the predictor and the expected survival times conditional on the predictor. Note that the estimator implemented in UnoC is restricted to situations where the random censoring assumption holds. The estimate for the C-index is:

```r
Cstat = UnoC(Surv.rsp, Surv.rsp.new, lpnew)
Cstat
```

```
## [1] 0.4761905
```